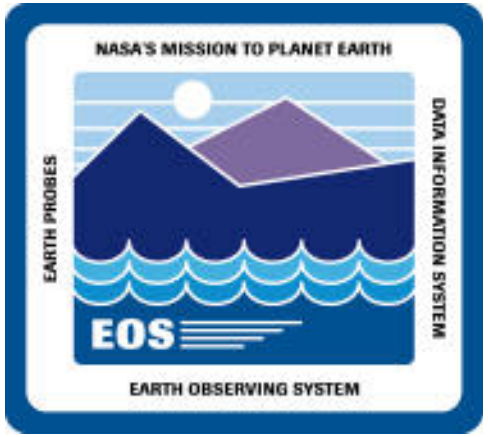


Detailed Data Server Subsystem Hardware Design

Alla Lake

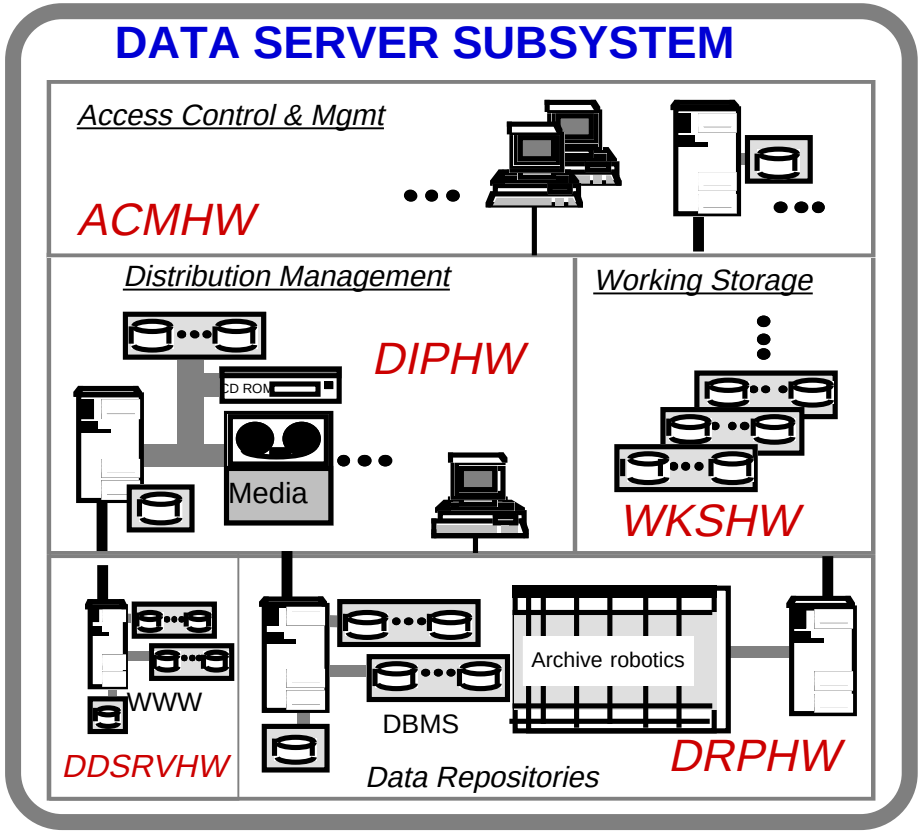
alake@eos.hitc.com

19 April 1996

Detailed Data Server Subsystem Hardware Design

1. Data Server Subsystem Configuration
2. Release B Hardware Design Drivers and Constraints
3. Design Methodology
4. Equipment Allocation (e.g., EDC)
5. Recoverability
6. Scalability
7. Conclusion

Data Server Subsystem Configuration



Detailed Data Server Subsystem Hardware Design

1. Data Server Subsystem Configuration
2. **Release B Hardware Design Drivers and Constraints**
3. Design Methodology
4. Equipment Allocation (e.g., EDC)
5. Recoverability
6. Scalability
7. Conclusion

Release B Hardware Design Drivers and Constraints

1. Data Rates
2. Volume Accumulations
3. Interim Products' Volume and Rates
4. Transaction Patterns
5. Archival Media Considerations

Design Drivers - Steady State Data Rates* (in MB/sec)**

Site	Total Input to th Archive	Total Outputs from the Archive
GSFC	12.19	39.39
LaRC	8.61	6.48
EDC	10.93	9.98
NSIDC	1.18	5.84
JPL	0.71	1.81
ASF	0.98	1.37

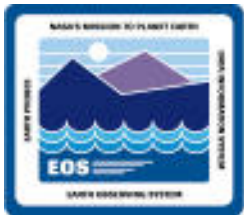
*** Static Analysis of the Technical Baseline**
**** Rates represent 1x Processing and 1x Reprocessing**

Design Drivers - Volume Accumulation*

Site	Epoch k (June 1999) [TB]	Year 2003 [TB]
GSFC	179.30	2,206.00
LaRC	49.44	485.00
EDC	194.64	1,061.00
NSIDC	6.02	40.70
JPL	7.46	18.80
ASF	1.11	2.50

*** Nominal Accumulations**

Design Drivers - Interim Products



Interim Products require non-permanent storage

Interim Products are intermediate products used in generating final, permanently archived, products

The main component of this accumulation of data is in support MODIS L3 processing (all the Interims produced between subsequent L3 processing campaigns)

Maximum lifetime of an Interim Product is 90 days

Design Drivers - Interim Products Steady State Data Rates* (in MB/sec)

Site	Interim Products Input to WKS	Interim Products Output from WKS
GSFC	11.20	2.60
LaRC	7.80	7.20
EDC	19.00	41.00
NSIDC	0.20	negligible
JPL	negligible	negligible
ASF	negligible	negligible

*** Static Analysis of the Technical Baseline**

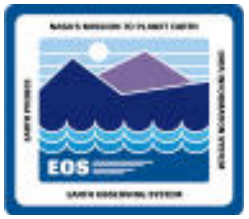
Design Drivers - Interim Products

Site	Peak Accumulation [GB]
GSFC	158.50
LaRC	314.80
EDC	6403.00
NSIDC	negligible
JPL	0.71

Impact:

- **At LaRC and GSFC additional quantities of disk are required at the WKS level**
- **At EDC high access rate ATLs (STK Powderhorns with Timberline linear drives) are added at the WKS level**

Design Drivers - Transaction Patterns (Files/Hour)*



Average Data File Size of 95 MB

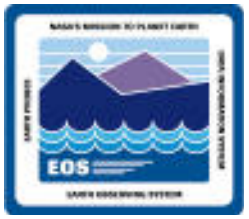
SITE	TOTAL INPUTS TO THE ARCHIVE	TOTAL OUTPUTS FROM THE ARCHIVE	INTERIM PRODUCTS INPUT TO WKS	INTERIM PRODUCTS OUTPUT FROM WKS
EDC	414	378	720	1554
LaRC	326	246	296	273
GSFC	462	1493	424	100
NSIDC	45	221	8	negligible
JPL	27	69	negligible	negligible
ASF	37	52	negligible	negligible

*** Average number of 95 MB Data File Transactions per hour based on 24 hour steady state rates**

Impact:

- **Use of faster Robotic devices (STK) at GSFC and EDC in combination with the Helical Scan D3 tape drives (due to high volume accumulation)**

Design Drivers - Archival Media Considerations



DSS is transaction bound at all sites

- **High performance linear scan tape drives (10 GB) (except at EDC and GSFC)**
- **EMASS Robotics (except at EDC and GSFC)**

DSS is also capacity driven at GSFC & EDC

- **High performance helical scan tape drives (50 GB)**
- **STK Robotics**

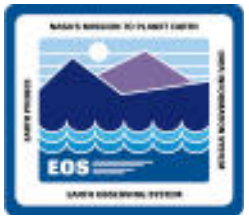
Detailed Data Server Subsystem Hardware Design

1. Data Server Subsystem Configuration
2. Release B Hardware Design Drivers and Constraints
3. **Design Methodology**
4. Equipment Allocation (e.g., EDC)
5. Recoverability
6. Scalability
7. Conclusion

Hardware Design - Methodology

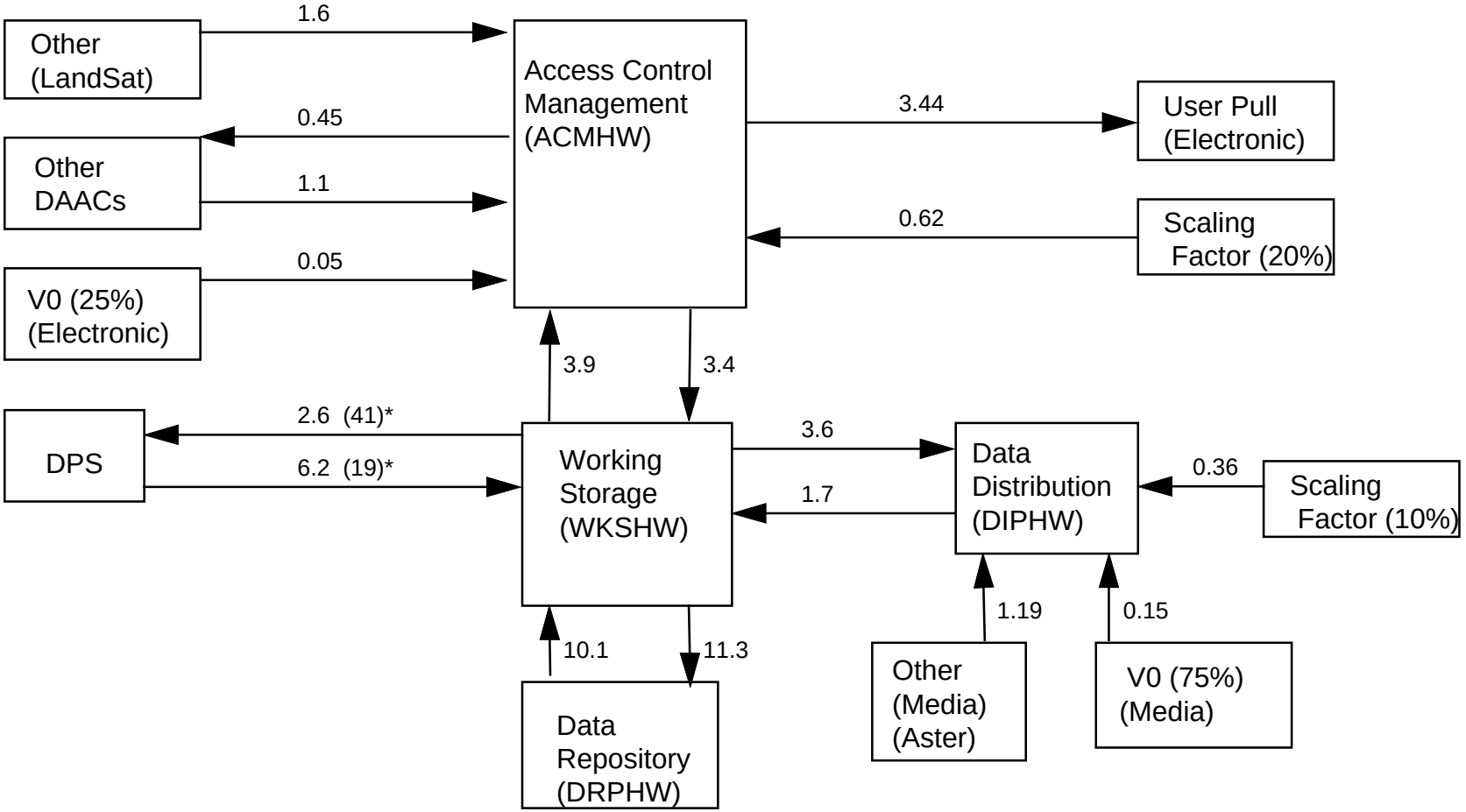
- **Evolutionary Process from Release A**
- **Steady State Modeling was Used to Determine the Hardware Design of the Data Repository and Working Storage**
- **Static Modeling of the February 1996 Technical Baseline by the Modeling Group Supplied Processing Data Flows and Interim Data Flows**
- **Applied Dynamic Modeling Results from the August 1995 Technical Baseline to Establish Transaction Patterns and Specific Data Flow Multipliers**

Hardware Design - Methodology



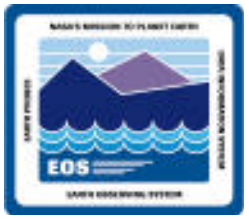
- **Data Rates**
- **Transaction Mitigation**
  - **Serendipity Factor**
- **Data Accumulations**

EDC Design Methodology - Data Rate Characterization (MB/s)



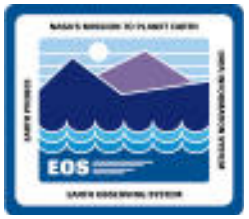
All Flows Represent Nominal Rates at Steady State over 24 hours in MB/s for the Technical Baseline (Epoch K) with some static inputs
* Interim Products

Data Repository Volume Groups



- **Volume groups are used in the Data Repository to spread data transfers among various tape resources.**
- **Increasing the number of volume groups per ATL, increases the number of concurrent writes possible per ATL.**
- **Spreading volume groups over multiple ATLs, provides fault tolerance to a system by insuring that not all data for a particular application will be in a single ATL.**

Hardware Design - Methodology (Host Sizing)



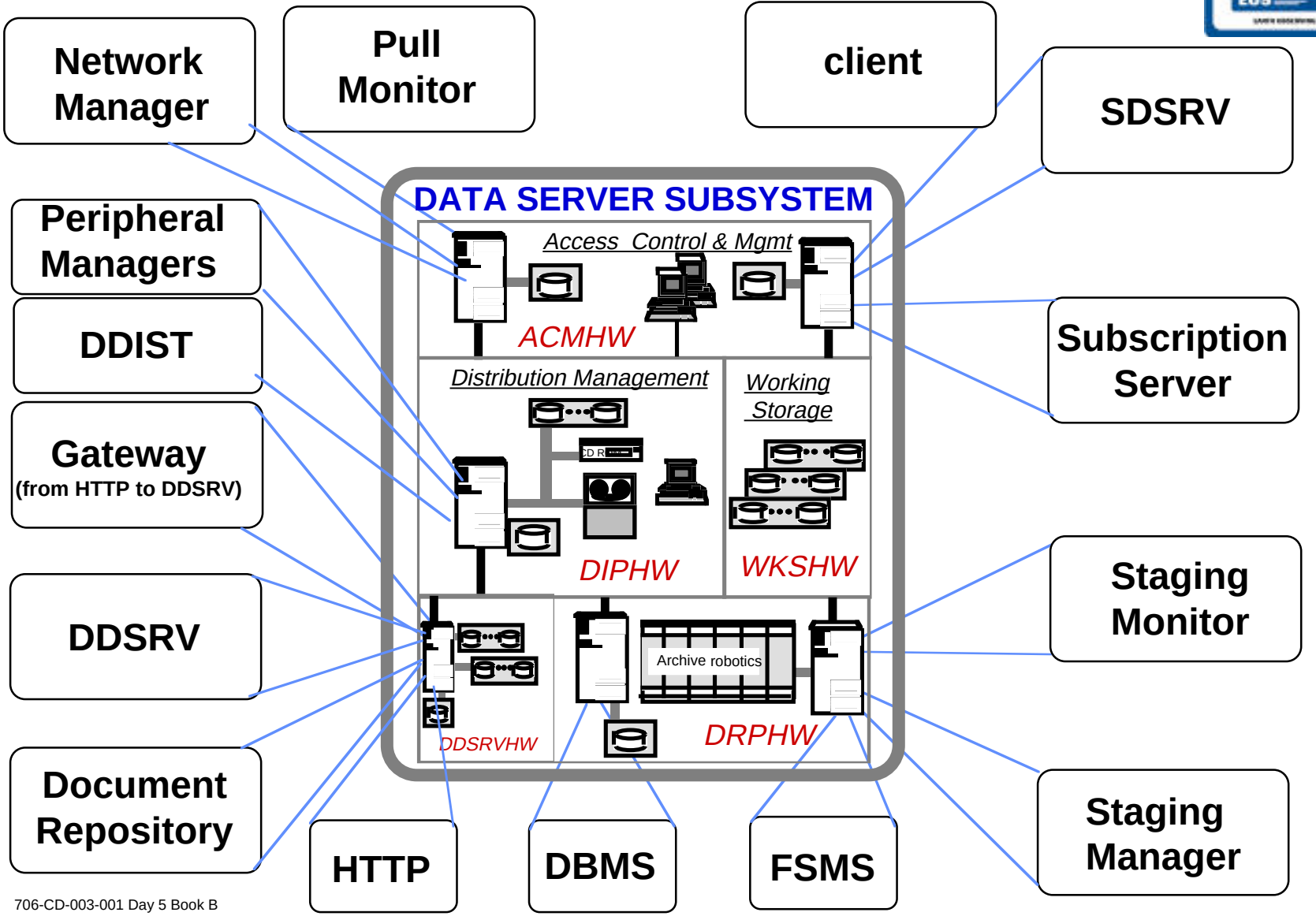
CPU Sizing

- **Processes**
- **Data Flow**
- **Peripheral Interfaces**
- **Communications interfaces**

Disk Sizing

- **Data Rates**
- **Data Residency**
- **FSMS Caching**
- **All RAID is specified as SGI RAID 5 (SCSI-2 based)**

Process Mapping



Host Sizing - Rationale

Quantifying CPUs for the hosts in each HWCI was accomplished using:

- **assumptions,**
- **testing,**
- **dynamic modeling outputs,**
- **SGI Rules of Thumb (SGI/RT),**
- **and combinations of two or more methods.**

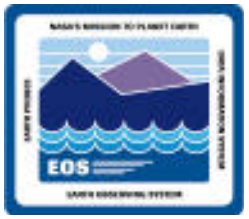
RAM allocation was determined using SGI/RT.

Host Sizing - Rationale (cont.)

The following characteristics were used for sizing (where appropriate):

1.	Subsetting / Subsampling	Dynamic Modeling
2.	Subscription Processing	Dynamic Modeling
3.	Monitoring Functions	Analysis
4.	Operating System / Applications	Analysis
5.	Clients (Session Threads)	Analysis
6.	HIPPI I/O	Analysis - Testing
7.	FDDI I/O	SGI/RT
8.	Total SCSI I/O	SGI/RT
9.	Total RAID I/O	SGI/RT
10.	FSMS	SGI/RT - Testing

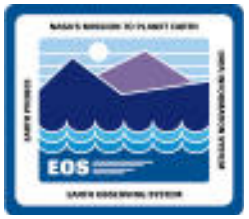
Why Use RAID?



Recoverability from a Single Disk Failure

- **RAID 5 can sustain the failure of one disk per group without data loss**
- **With a single disk failure, the array stays on line with degraded performance**
- **The SGI RAID units can replace/rebuild a single failed disk automatically using a hot spare**
- **At Epoch K, we will have approximately 221 drives in DSS at EDC**
  - **Predicted MTBF per drive is ~300,000 hours**
  - **On the average, at EDC, a drive failure will occur every 2 months**
  - **With striped filesystems required for Release B performance, loss of a single drive may require rebuilding a filesystem consisting of 16 (or more) drives (68 to 144 GB)**

Why Use RAID? (cont.)



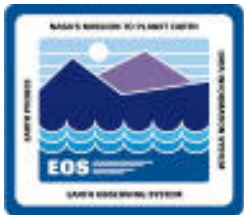
Without RAID, Limited to 7 Disks per SCSI-2 Channel

- **Most systems currently configured with 8 to 16 data disks per channel; would need ~20% more SCSI-2 channels, totaled across all DSS systems**
- **Number of SCSI-2 HIO and IO4 boards would increase**

The ECS Cost Ratio for RAID 3 or 5 to Non-RAID Is 1.37:1.00

- **Based on a quotation for 500 GB of net storage for both approaches received from SGI on 2/28/96**
- **Does not take into account additional costs of non-RAID solution for additional IO4 and SCSI-2 HIO cards**

Hardware Design - Methodology (Peripheral Sizing)



Automated Tape Library Sizing

- Data Rates
- Transaction Rates
- Volume Accumulation

Distribution Peripheral Sizing

- Data Rates
- Operational Hours

Detailed Data Server Subsystem Hardware Design

1. Data Server Subsystem Configuration
2. Release B Hardware Design Drivers and Constraints
3. Design Methodology
4. **Equipment Allocation (e.g., EDC)**
5. Recoverability
6. Scalability
7. Conclusion

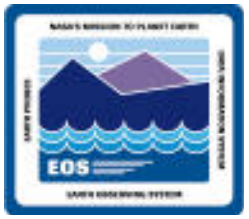
Data Server Subsystem
EDC Hardware

HW CI	Equipment	Quantity
ACMHW		
Admin. Workstations	SUN Ultra	2 ea.
APC Hosts	22 CPU SGI Challenge XL, 6 GB local disk, 2 GB RAM	1 ea.
	12 CPU SGI Challenge XL, 6GB local disk, 1 GB RAM	1 ea.
APC Disk	682.5 GB of RAID Disk	
WKSHW		
	10 CPU SGI Challenge XL, 6 GB local disk, 1 GB RAM	1 ea.
	6 CPU SGI Challenge XL, 6 GB local disk, 512 MB RAM	1 ea.
	421 GB RAID	
	STK Powderhorn ATLs with 16 3490e tape drives	2 ea.
DRPHW		
FSMS Server Host	6 CPU SGI Challenge XL, 6 GB local disk, 512 MB RAM	2 ea.
AMASS Cache	50 GB RAID Disk for AMASS cache (total of 300 GB RAID)	
ATL Robotics	STK Powderhorn with 16 D3 drives per ATL	2 ea.
Tape Media	D3 50 GB cartridges	8,042 ea.
DBMS Server	2 CPU SGI Challenge XL, with 20 GB shared disk	2 ea.
DIPHW		
Staging Server Host	Sun Ultra 4-slots, with access to 310GB RAID disk	2 ea.
Peripherals:	8 mm tape drives (with stackers)	12 ea.
	4mm tape drives (with stackers)	12 ea.
	6250 tape drive	1 ea.
	CD ROM drive and jukebox	20 ea.
	FAX	1 ea.
	3480/3490 outboard drives	4 ea.
	printer	2 ea.
DDSRVHW		
WAIS/http Data Server	CPU SMP Server - SUN Ultra 4-slot, 6 GB local disk, 256 MB RAM	2 ea.
Data Server Disk	6 GB mirrored in two machines	

Detailed Data Server Subsystem Hardware Design

1. Data Server Subsystem Configuration
2. Release B Hardware Design Drivers and Constraints
3. Design Methodology
4. Equipment Allocation (e.g., EDC)
5. **Recoverability**
6. Scalability
7. Conclusion

Recoverability



Static Modeling Projection of System Recoverability

- **Recovery within the 24 hours period of failure**
- **Mean Down Time = 2 Hours (Level 3 Requirement)**
- **Max Down Time = 4 Hours (Level 3 Requirement)**
- **Recover via limiting reprocessing and/or user pull service until fully recovered**

Recovery Scenario (Robot Arm Failure)

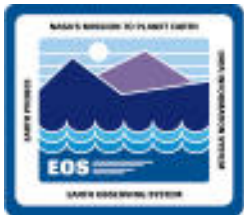
- **Degraded Mode Operations at sites with more than one robot**
- **Distribution of data flow among two or more robots**
- **Recovery within 24 hours**

- **Down Time at the sites with a single robot (ASF, JPL)**
- **Recovery within 24 hours**

Detailed Data Server Subsystem Hardware Design

1. Data Server Subsystem Configuration
2. Release B Hardware Design Drivers and Constraints
3. Design Methodology
4. Equipment Allocation (e.g. EDC)
5. Recoverability
6. **Scalability**
7. Conclusion

DSS Transaction Rates Capacity Margin



The DSS Transaction Rates Capacity Margin reflects the additional I/O, Processing and Storage excess capacity at the following sites.

EDC is ~ 5%
LaRC is ~ 60%
GSFC is ~ 5%
NSIDC is > 20%
JPL is > 20%
ASF is > 20%
ORNL is > 20%

~ Approximately
> Greater Than

Scalability Breakpoints - 4X Distribution

Delta for EDC:

1 ea	-	STK/D3,
9 ea	-	D3,
2 ea	-	STK/3490e,
19 ea	-	3490e,
1	-	Sun Ultra,
11	-	8 mm Drives,
11	-	4 mm Drives,
18	-	CD ROM.

Scalability Breakpoints - 10X Distribution

Delta for EDC:

2	-	STK/D3,
36	-	D3,
7	-	STK/3490e,
74	-	3490e,
2	-	Sun Ultra,
2	-	SGI XL,
42	-	8 mm Drives,
42	-	4 mm Drives,
71	-	CD ROM.

Detailed Data Server Subsystem Hardware Design

1. Data Server Subsystem Configuration
2. Release B Hardware Design Drivers and Constraints
3. Design Methodology
4. Equipment Allocation (e.g., EDC)
5. Recoverability
6. Scalability

7. Conclusion